TOWARDS TAMPER DETECTION AND CLASSIFICATION WITH ROBUST WATERMARKS

Henry Knowles, Dominique Winne, Nishan Canagarajah, David Bull

Image Communications Group, Centre for Communications Research, University of Bristol, Woodland Rd, Bristol BS8 1UB, UK

ABSTRACT

In this paper, we present a novel wavelet-based double watermarking system for the detection and subsequent characterisation of the tampering of images. Most tamper detection techniques use fragile watermarks. However, our previous work showed that this type of watermark is frequently completely destroyed by compression, which in many cases is undesirable. In addition, it gives little or no information about the nature of the attack. We propose using two robust watermarks, one inserted at the embedder, the other at the detector. The second watermark is used as a reference with which the first watermark may be compared. This allows additional information not previously available about the nature of the attack to be obtained. The use of a robust strategy prevents the watermark being easily destroyed, but instead allows the system to continue to perform after significant attacks.

1. INTRODUCTION

Initially, image authentication used fragile watermarks that were inserted either in the spatial domain [7], the wavelet domain [9] or the DCT domain [10]. Whilst these techniques frequently provide accurate tamper resolution, they do not permit the compression of the image using lossy compression techniques such as JPEG compression, nor do they provide information about the nature of the attack. A considerable amount of attention has been given to the problem of overcoming the compression issue. For example, Winne *et al.* [8] offer a system where compression up to a fixed Quality Factor is permissible by relaxing the tamper detection criterion used. Ideally, some measure invariant to compression would be watermarked, such that the watermark would be unaffected by the compression, but destroyed by another form of attack. Lin *et al.* [2, 3] use the relationship between DCT coefficients before and after compression to authenticate images, whilst various authors have considered using compression invariant features such as the moments of watermark blocks [5] and the use of feature points which are essential to the images semantic content, but should be unaffected by compression [1]. However, both of these techniques require additional information to be stored, which is in many situations unacceptable. Additionally, none of these methods provide information about the nature of the attack. We therefore propose an algorithm which uses two robust watermarks, one inserted at the embedder, and the other at the detector, to derive additional information about the nature of any attack to which the image has been subjected.

In the next section (Section 2), we present details of how our algorithm works and a justification for the method. This is followed by Section 3, where we present some experimental results to demonstrate the potential of the algorithm. In Section 4 conclusions are presented, along with some ideas for future work.

2. ALGORITHM DESCRIPTION

2.1. Watermark Embedding

Watermark insertion is performed in the wavelet domain using the Noise Visibility Function (NVF) based model of Voloshynovskiy *et al.* [6]. The NVF is calculated in the wavelet domain on a subband-by-subband basis as:

$$NVF(i,j) = \frac{1}{1 + \theta \,\text{Local Variance}(i,j)} \quad (1)$$

where $\theta = D/\text{Max Variance}, 50 < D < 100$. Thus the watermark strength, $S(i,j)$ is given by:

$$S(i,j) = S_{e_{l,o}}(1 - NVF(i,j)) + S_{f_{l,o}} NVF(i,j) \quad (2)$$

where $S_{e_{l,o}}$ represents the strength for the edge regions in decomposition level l and orientation o, and $S_{f_{l,o}}$ is a similar measure for the flat regions. Finally, the watermarked image coefficient $I'(i,j)$, may be defined as:

$$I'(i,j) = I(i,j) + w(i,j)S(i,j) \quad (3)$$

where $I(i,j)$ is the original image coefficient, and $w(i,j)$ is the watermark wavelet coefficient. The watermark is tiled over the image in blocks of 32-by-32 pixels and is defined as $w(i,j) \in [-1,1]$. The purpose of the tiling is to enable localisation of the attacks, whilst the block size is set such that the watermark will be more robust to attacks, without becoming so large that the localisation is not meaningful.

2.2. Watermark Estimation

The watermark is extracted using wavelet denoising techniques as given in Moulin *et al.* [4]. An estimate of the watermark (or noise) -free image is determined using the thresholding process. This estimate is then subtracted from the received, watermarked and possibly attacked image to give an estimate of the watermark. Either soft or hard wavelet thresholding may be used. The measure used to judge the degradation of the watermark is the coherence or normalised correlation coefficient, ρ:

$$\rho(X,Y) = \frac{R_{XY}}{\sqrt{R_{XX}R_{YY}}} \quad (4)$$

where R denotes the variance or covariance. Thus for two identical signals (invariant to scaling), the coherence will be 1, whereas for

0-7803-7761-3/03/$17.00 ©2003 IEEE

two uncorrelated signals, the coherence will be 0. Negative coherence implies that the signals are correlated but have been shifted in phase by π rads. In order to maximise the potential of the algorithm to work with as severe an attack as possible, the fact that the watermark is fully defined at the detector is used. Instead of blindly estimating the watermark and image variances as is done for robust watermarks [4, 6], a selection of different thresholds are tried, and the one that maximises the coherence between the estimated watermark and the original is chosen. The range of thresholds used will vary between the minimum and maximum wavelet coefficients in the current watermark block. The same threshold is used for all coefficients in a given watermark block. Thus the watermark extraction process may be defined as:

$$\hat{w} = \max_{T}[\text{Coherence}(\{I'' - \text{Threshold}(I'', T)\}, w)] \quad (5)$$

2.3. Algorithm Overview

Results from preliminary experiments showed that firstly the values obtained for ρ, the coherence of the estimated watermark with the original watermark, varied considerably over a given subband. Different attacks caused a similar variation. The second finding was that NVF of the untampered watermarked image was very similar to the NVF of the original image. From the first observation, there is clearly a need for a reference to decide whether ρ has assumed a particular value due to an attack or the statistical content of the watermark block. The second observation provides an insight into what the solution might be. By adding a second watermark to the image at the receiver and then immediately estimating it, a reference to what the first watermark would have been had there been no attack is obtained.

In summary, the primary watermark (w_1) is embedded (see Section 2.1) in the original image I, to form the watermarked image, I'. The watermarked image is stored, and whilst in storage may or may not be attacked. An estimate of the primary watermark ($\hat{w}_1$) is generated as described in Section 2.2 from the possibly attacked, watermarked image I''. The coherence (see Equation 4) of w_1 with $\hat{w}_1$ is denoted ρ_1. The second watermark, w_2 is now added to I'' and immediately estimated as before to give $\hat{w}_2$. Again the coherence is calculated between w_2 and $\hat{w}_2$ to give ρ_2. By comparing ρ_1 and ρ_2, information about the nature of the attack can be gained. It is important to note that w_1 and w_2 are chosen to be uncorrelated with each other.

3. RESULTS

Results for a variety of different attacks are presented in this section. The attacks considered are: no attack, compression using the JPEG2000 algorithm with a compression ratio of 100:1, compression using the JPEG baseline with a quality factor of 80, and an unsharp mask as found in the MATLAB Image Processing Toolbox.

Results for ρ_1 are given in Figure 1 and also summarised in Table 1. It can be seen that there is considerable variation in ρ_1 across the subbands. However, other more positive trends are also visible. A comparison of Figure 1(a) with 1(b) (or Table 1(a) with 1(b)) shows that the JPEG2000 compression has severely affected the watermark, and thus the value obtained for ρ_1 is very low in the high frequency subbands (about 0.1), whilst the value for the untampered case is much higher; about 0.6 for the level 1 subbands. The unsharp mask (Figure 1(c)) reduces the watermark coherence

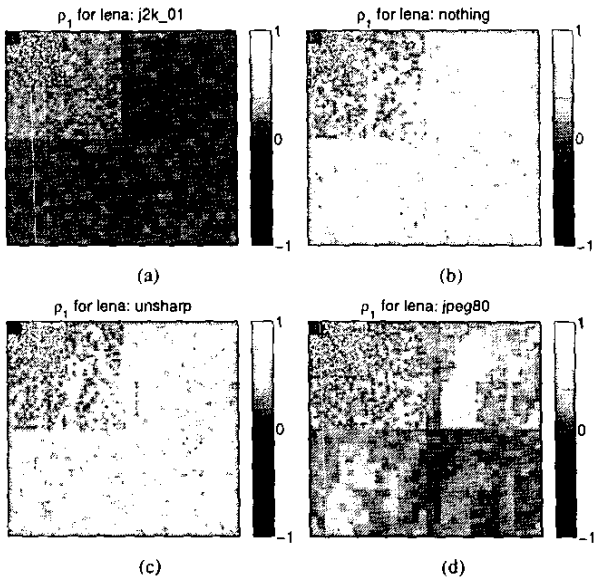

Figure 1: ρ_1 for Lena for attacks: (a) Compression using JPEG2000 with compression ratio of 100:1, (b) No attack, (c) Unsharp mask, and (d) Compression using JPEG baseline with a quality factor of 80

only very slightly, whilst the JPEG baseline attack (Figure 1(d)) is in between the unsharp and JPEG2000 attacks. These observations confirm what might have been intuitively expected. Compressing an image using JPEG2000 with a compression ratio of 100:1 will result in a significant loss of information, especially in the low scale/high frequency subbands, and this is as observed. The JPEG baseline compression is less severe, and so will result in a lower, but still visible change, and again most of the information lost will be in the higher frequency subbands. An unsharp mask however, has the effect of amplifying the high frequency subbands. As the coherence is a linear function and invariant to variance changes in X and Y (see Equation 4), one would not expect ρ_1 to change significantly. For example, attacking a particular subband by multiplying it by a constant will not affect the value of ρ_1. What will affect ρ_1 is the rounding and clipping that will take place when the image is converted back into the pixel domain. Low pass filtering, for example, may alter the image such that after rounding to integer values in the pixel domain, the high frequency wavelet coefficients are all, or nearly all, zero. Thus ρ_1 will be affected. However, the unsharp mask is unlikely to result in much clipping taking place (e.g. causing pixel values to lie outside the range 0-255), so as was stated earlier the effects of this attack on ρ_1 will not be great. This may explain why the changes to ρ_1 are small when compared to the changes in PSNR as given in Table 3.

If the results in Figure 2 and Table 2 are now examined, additional information can now be obtained. It can be seen that those attacks which removed high frequency information (Figures 2(a) and 2(d)) have high values for ρ_2, and indeed the more information that was removed, the higher ρ_2 becomes. In comparison, the unsharp mask, which amplified the high frequency information, has a lower value for ρ_2. This effect is particularly clear in the Level 1 & 2 diagonal subbands (compare Table 1(c) with Table 2(c)). The

Table 1: Table showing ρ_1 at different decomposition levels and in different subbands for Lena after attack (a) J2K r=0.01, (b) nothing, (c) Unsharp mask, and (d) JPEG compression with QF = 80

(a)

	Orientation		
	H	V	D
L1	0.10	0.11	0.11
L2	0.21	0.23	0.20
L3	0.42	0.35	0.36
L4	0.50	0.54	0.49

(b)

	Orientation		
	H	V	D
L1	0.59	0.57	0.58
L2	0.41	0.37	0.55
L3	0.61	0.45	0.52
L4	0.53	0.56	0.53

(c)

	Orientation		
	H	V	D
L1	0.54	0.53	0.55
L2	0.40	0.38	0.50
L3	0.61	0.47	0.53
L4	0.54	0.58	0.54

(d)

	Orientation		
	H	V	D
L1	0.34	0.30	0.16
L2	0.35	0.33	0.45
L3	0.60	0.44	0.49
L4	0.53	0.56	0.53

Table 2: Table showing ρ_2 at different decomposition levels and in different subbands for Lena after attack (a) J2K r=0.01, (b) nothing, (c) Unsharp mask, and (d) JPEG compression with QF = 80

(a)

	Orientation		
	H	V	D
L1	0.39	0.37	0.44
L2	0.59	0.52	0.90
L3	0.62	0.62	0.80
L4	0.66	0.55	0.75

(b)

	Orientation		
	H	V	D
L1	0.63	0.62	0.66
L2	0.40	0.35	0.57
L3	0.50	0.49	0.64
L4	0.53	0.45	0.65

(c)

	Orientation		
	H	V	D
L1	0.54	0.48	0.36
L2	0.37	0.34	0.43
L3	0.48	0.47	0.62
L4	0.53	0.44	0.63

(d)

	Orientation		
	H	V	D
L1	0.54	0.53	0.53
L2	0.40	0.36	0.57
L3	0.51	0.49	0.63
L4	0.52	0.46	0.65

value of ρ_2 for the untampered case is largely unchanged. These observations may be explained heuristically as follows. As a result of the attacks, the amount of energy in the subbands changes, especially, for the attacks in question, in the high frequency subbands. However, the masking function (see Equation 1) is nonlinear with energy (variance). Therefore it changes at a different rate to the energy. In fact, it changes more slowly than the energy does. Therefore, the NVF calculated at the detector will be more similar to the NVF calculated at the embedder than the respective image energies at the detector and the embedder. Thus the ratio of watermark to image energy will change; it will increase for low pass attacks, and decrease for the unsharp mask. This in turn will mean that the thresholding process will be more and less able, respectively, to better separate the watermark from the image.

Figure 3 gives results showing how the average watermark coherence changes between different wavelet decomposition levels for a number of different images e.g. the average of $\rho_1 - \rho_2$. It can be seen for a given decomposition level and attack, the variation between images is generally small. In addition, it should also be noted that the profiles for the different attacks are quite different. For example, the results for the unsharp mask (Figure 3(c)) are higher for the diagonal subbands in levels 1 and 2 than for the untampered case (Figure 3(b)). The frequency response of the unsharp filter has its largest magnitude in the high frequency corners of the Fourier domain, thus one would expect it to most affect the diagonal components of the image. This correlates with what was observed previously. It may also be noted that the average difference in coherence for the untampered case is close to zero for all the high frequency subbands, which is what one might hope for. Conversely, the average change for either type of compression (see Figures 3(d) and 3(a) for JPEG baseline and JPEG2000 compression respectively) is generally negative, rather then positive as was the case for the unsharp mask. This is also as expected. Interestingly, JPEG2000 also has a noticeable effect on wavelet subband levels 3 & 4, which none of the other attacks do. The fluctuation in all the results in the level 3 & 4 subbands is due to the fact that

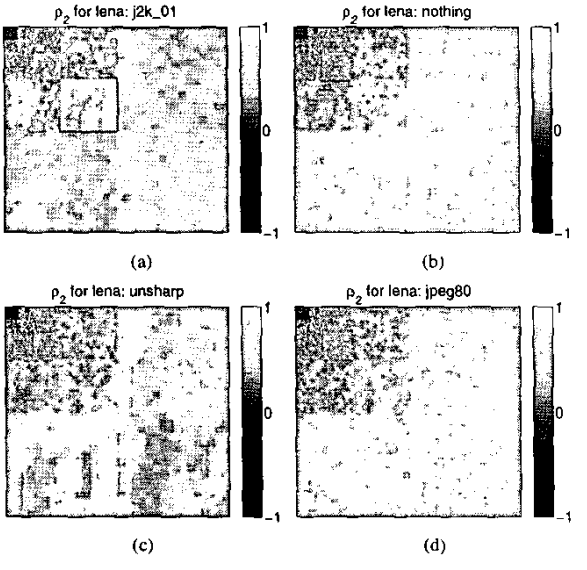

Figure 2: ρ_2 for Lena for attacks: (a) Compression using JPEG2000 with compression ratio of 100:1, (b) No attack, (c) Unsharp mask, and (d) Compression using JPEG baseline with a quality factor of 80

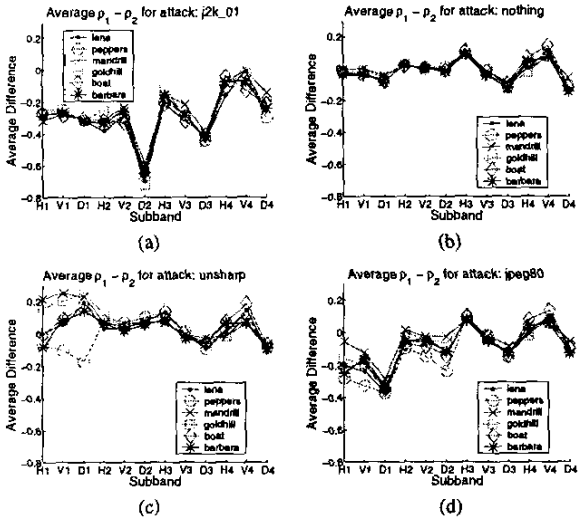

Figure 3: Average difference between ρ_1 and ρ_2 for attacks: (a) Compression using JPEG2000 with compression ratio of 100:1, (b) No attack, (c) Unsharp mask, and (d) Compression using JPEG baseline with a quality factor of 80

Table 3: Table showing PSNR (dB) for various attacks compared to watermarked untampered version of Lena

Attack	JPEG QF=80	J2K r = 0.01	Unsharp
PSNR(dB)	35.61	27.67	22.44

natural images tend to have more energy in the low frequency subbands, which is also the area in which the Human Visual System is most sensitive. Therefore less watermark energy can be embedded in these regions, with the result that the thresholding process will be less able to separate the image and the watermark. This in turn will affect ρ_1 and ρ_2, making them more image dependent and thus less predictable.

4. CONCLUSIONS

In this paper, we have presented a novel double-watermarking algorithm that uses a second reference watermark against which the first watermark may be compared to determine not only whether the image has been attacked, but also information about the nature of the attack. We have shown that a variety of different images have similar metrics for the same attack, thus suggesting that these or other metrics may be used to discriminate between attacks. Additional conclusions are that the use of the double watermarking strategy is best suited to the wavelet domain, or indeed any other domain with both spatial and frequency resolution, as this allows comparisons over different subbands. This, as shown, permits greater distinctions between the attacks to be observed.

In the future, there needs to be developed some method of classifying different attacks, which must also provide better localisation of the attacks. Also, a larger number of attacks should be considered; in particular, the case of replacement needs to be addressed. Additional metrics may also need to be developed to allow more accurate separation of the attacks, especially when comparing the unsharp mask and the untampered case. More generally, special attention may need to be directed towards the case of the linear attack, as it has been shown that this can be more difficult to detect than non-linear operations.

5. ACKNOWLEDGMENTS

This work is supported by Motorola, the Metropolitan Police and EPSRC grant GM/M81885.

6. REFERENCES

[1] M. Kutter, S. K. Bhattacharjee, and T. Ebrahimi. Towards second generation watermarking schemes. In *Proceedings ICIP-99 (IEEE International Conference on Image Processing)*, volume 1, October 25–28 1999.

[2] C-Y Lin and S-F Chang. Robust image authentication method surviving JPEG lossy compression. In *SPIE Storage and Retrieval of Image/Video Database*, volume 3312, 1998. San Jose.

[3] C-Y Lin and S-F Chang. SARI: Self-Authentication-and-Recovery Image watermarking system. In *ACM Multimedia*, September 30–October 5 2001.

[4] P. Moulin and J. Liu. Analysis of multiresolution image denoising schemes using generalised-gaussian and complexity priors. *IEEE Transactions on Information Theory Special Issue on Multiscale Analysis*, April 1999.

[5] M. P. Queluz. Authentication of digital images and video: Generic models and a new contribution. *Signal Processing: Image Communication*, 16:461–475, 2000.

[6] S. Voloshynovskiy, F. Deguillaume, and T. Pun. Content adaptive watermarking based on a stochastic multiresolution image modeling. In *Tenth European Signal Processing Conference (EUSIPCO'2000)*, September 5–8 2000.

[7] S. Walton. Information authentication for a slippery new age. *Dr. Dobbs Journal*, 20(4):18–26, April 1995.

[8] D. A. Winne, H. Knowles, D. R. Bull, and C. N. Canagarajah. Compression compatible digital watermark algorithm for authenticity verification and localization. In *Security and Watermarking of Multimedia Contents IV*, volume 4675. SPIE, January 2002.

[9] D. A. Winne, H. Knowles, D. R. Bull, and C. N. Canagarajah. Digital watermarking in wavelet domain with predistortion for authenticity verification and localization. In *Security and Watermarking of Multimedia Contents IV*, volume 4675. SPIE, January 2002.

[10] M. Wu and B. Liu. Watermarking for image authentication. In *Proceedings ICIP-98 (IEEE International Conference on Image Processing)*, 1998.